Membership Inference Attacks on Machine Learning: A Survey

Year: 2021 | Citations: 561 | Authors: Hongsheng Hu, Z. Salcic, Lichao Sun, G. Dobbie, P. Yu

Abstract

Machine learning (ML) models have been widely applied to various applications, including image classification, text generation, audio recognition, and graph data analysis. However, recent studies have shown that ML models are vulnerable to membership inference attacks (MIAs), which aim to infer whether a data record was used to train a target model or not. MIAs on ML models can directly lead to a privacy breach. For example, via identifying the fact that a clinical record that has been used to train a model associated with a certain disease, an attacker can infer that the owner of the clinical record has the disease with a high chance. In recent years, MIAs have been shown to be effective on various ML models, e.g., classification models and generative models. Meanwhile, many defense methods have been proposed to mitigate MIAs. Although MIAs on ML models form a newly emerging and rapidly growing research area, there has been no systematic survey on this topic yet. In this article, we conduct the first comprehensive survey on membership inference attacks and defenses. We provide the taxonomies for both attacks and defenses, based on their characterizations, and discuss their pros and cons. Based on the limitations and gaps identified in this survey, we point out several promising future research directions to inspire the researchers who wish to follow this area. This survey not only serves as a reference for the research community but also provides a clear description for researchers outside this research domain. To further help the researchers, we have created an online resource repository, which we will keep updated with future relevant work. Interested readers can find the repository at https://github.com/HongshengHu/membership-inference-machine-learning-literature.